Homework Assignment 5

Due date: Thursday, December 10 (hardcopy in class)

1. In this question we consider the algorithm for learning conjunctions as discussed in class (see also section 7.3.2 in Mitchell’s textbook). What happens if we run this algorithm on a noisy dataset where no conjunction is consistent with the data? Illustrate your answer with an example dataset.

2. Consider learning classifiers in a problem where each example is a two dimensional point with integer values and each classifier is given by a rectangle on the grid where the grid is bounded between \((0,0)\) and \((M,M)\) where \(M\) is a large number. Every classifier is given by its bottom-left point and top-right point and classifies all the points inside the rectangle as positive and points outside as negative. For example the rectangle \(((1,17),(5,55))\) classifies the points \((1,12)\) and \((6,75)\) as negative and the point \((3,22)\) as positive.

   (a) Give a simple upper bound on the number of classifiers in this class by considering the number of bottom-left and top-right points possible.

   (b) You are given 4 labeled examples: \((2,7)\) is negative, \((10,93)\) is positive, \((82,33)\) is positive, \((84,55)\) is negative. What concepts are consistent with this dataset?

   (c) Considering the PAC analysis given by Eq (7.2) in Mitchell’s textbook: How many examples are needed to guarantee with probability \((1-\delta) = 0.99\) that the error of a consistent learner is at most \(\epsilon = 0.01\)? Please explain your answer.

   (d) What is the bound if we change \(\epsilon\) to be 0.001? and what is the bound if we change \(\delta\) to be 0.001?

   (e) Can you give an efficient algorithm (polynomial in the number of examples; independent of \(M\)) that finds a consistent concept and can therefore be used with such guarantees?

   (f) Recall that in the agnostic PAC learning model we are not guaranteed that a consistent hypothesis exists. How many examples are needed for the same problem but for the agnostic PAC analysis given by Eq (7.3) in the textbook for \(\epsilon = \delta = 0.01\)?

   (g) Can you give a simple (but less efficient) algorithm for this case (polynomial in the number of examples and \(M\))?  

3. In this question we consider the primal perceptron and dual perceptron algorithms and how they work given a concrete dataset and kernel. For the algorithms please refer to the pseudocode in programming project 4. In particular consider the dataset below where we have 3 features and 2 examples, and consider the kernel \(k(x,y) = (x \cdot y)^2\).

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.33</td>
<td>0.78</td>
<td>1.02</td>
<td>(−)</td>
</tr>
<tr>
<td>0.44</td>
<td>1.25</td>
<td>0.88</td>
<td>(+)</td>
</tr>
</tbody>
</table>

   (a) Note that because the algorithms initialize their weight vector to 0 they will predict \((+)\) on the first example and therefore make a mistake. What is the hypothesis of the Primal Perceptron and Dual perceptron (vector \(w\) or \(\alpha\) respectively) after this first mistake? Please show any computation and explain your answer.

   (b) Now use the answer to the previous part to show the prediction of the algorithms on the second example, and if the algorithm makes a mistake show its hypothesis after the corresponding update.

   (c) For the dual perceptron show the expanded feature vector corresponding to the kernel and repeat the previous two parts, showing the results in terms of the sector \(w\) in the expanded space (you should get the same sums and predictions).
4. Recall that in class we have shown that \( k(x, y) = (x \cdot y + 1)^2 \) is a kernel. Now consider a new function \( f(x, y) = 4k(x, y) + 25 \). Prove that \( f() \) is a kernel by showing that it is an inner product, i.e., \( f(x, y) = \phi(x) \cdot \phi(y) \) for some \( \phi() \).

Note that you do not need to repeat the proof from class, but to use the fact that \( k(x, y) \) is an inner product for some space to prove the same for \( f(x, y) \).

5. In this question we consider the Naive Bayes algorithm and active learning. Consider the dataset below where the first 6 examples are labeled and the last 4 are not. Consider active learning with uncertainty sampling (where in the notation of the slides from lecture we pick the instance with the smallest \( p_1 \)).

<table>
<thead>
<tr>
<th></th>
<th>000</th>
<th>011</th>
<th>011</th>
<th>110</th>
<th>101</th>
<th>100</th>
<th>010</th>
<th>101</th>
<th>110</th>
<th>001</th>
</tr>
</thead>
<tbody>
<tr>
<td>Label</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(a) Show the hypothesis of Naive Bayes after learning from the 6 labelled examples.

(b) Which example will be picked by uncertainty sampling? explain your answer